

### Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems

Gippert, Jost

Veröffentlichungsversion / Published Version  
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:  
GESIS - Leibniz-Institut für Sozialwissenschaften

#### Empfohlene Zitierung / Suggested Citation:

Gippert, J. (2002). Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems. *Historical Social Research*, 27(1), 207-214. <https://doi.org/10.12759/hsr.27.2002.1.207-214>

#### Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:  
<https://creativecommons.org/licenses/by/4.0/deed.de>

#### Terms of use:

This document is made available under a CC BY Licence (Attribution). For more Information see:  
<https://creativecommons.org/licenses/by/4.0>

## Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems\*

*Jost Gippert*

### 1. Das TITUS-Projekt

Das Projekt eines "Thesaurus indogermanischer Text- und Sprachmaterialien" wurde 1987 mit einem in der Zeitschrift "Die Sprache" veröffentlichten Aufruf initiiert, der darauf abzielte, die Bemühungen um eine digitale Erfassung von Textmaterialien in altindogermanischen Sprachen zu koordinieren und somit einen für die fachlich interessierte Öffentlichkeit gemeinsam zur Verfügung stehenden Datenpool zu schaffen. Das ehrgeizige Ziel, die textuale Bezeugung der alten Indogermania, die sich vom ältesten (vedischen) Altindischen über altiranische Sprachen wie das Avestische oder das Altpersische, altanatolische Sprachen wie das Hethitische oder das Luvische, die Sprachen des Klassischen Altertums (Griechisch und Latein) bis hin zu den Überlieferungen der germanischen und keltischen Völker (z.B. in Form des Altirischen) erstreckt, in elektronischer Form für die wissenschaftliche Analyse zugänglich zu machen, hat in den seither vergangenen knapp 15 Jahren mehr und mehr konkrete Konturen angenommen, wobei insbesondere die im Jahre 1995 erfolgte Errichtung eines eigenen Web-Servers, von dem bereits verfügbare Materialien abgerufen werden können, zu einer Intensivierung der gemeinschaftlichen Datenaufbereitung führte. Von den mehr als 150 Beiträgern, die sich teils in Europa, teils auch in

---

\* Aus dem Protokoll des 83. Kolloquiums über die Anwendung der Elektronischen Datenverarbeitung in den Geisteswissenschaften an der Universität Tübingen vom 17. November 2001.

Achtung: Dieser Text ist in der Vorlage mit Unicode / UTF8 kodiert. Um die in ihm erscheinenden Sonderzeichen auf Bildschirm und Drucker sichtbar zu machen, muß ein Font installiert sein, der Unicode abdeckt wie z.B. der TITUS-Font Titus Cyberbit Unicode. In der Folge sind bei Sonderzeichen nur Stellvertreter gedruckt. Zuerst veröffentlicht unter <http://www.uni-tuebingen.de/zdv/zrinfo/prot/prot831-titus.html>.

Übersee an dem Projekt beteiligt haben, ist ein Datenbestand erarbeitet worden, der inzwischen weit mehr als 1 GB reiner Textdaten umfasst, wobei durch vergleichbare Projekte abgedeckte Bereiche wie derjenige der altgriechischen Überlieferung, für den der "Thesaurus Linguae Graecae" der Universität von Kalifornien und das "Perseus"-Projekt der Tufts-University zuständig sind, weitgehend ausgeklammert bleiben.

Die somit umrissene Textdatenbank bildet nach wie vor das Kernstück des TITUS-Servers, der unter den Webadressen <http://titus.uni-frankfurt.de> und <http://titus.fkidg1.uni-frankfurt.de> zugänglich ist und von dem außer den Textmaterialien selbst verschiedene allgemeine Informationen aus dem Bereich der (indogermanischen und allgemeinen) Sprachwissenschaft, Lehrmaterialien wie Sprachenkarten oder Audioaufnahmen gesprochener Sprachen und anderes mehr abrufbar sind. Als Online-Textdatenbank will der TITUS-Thesaurus heute nicht mehr nur die Überlieferungen in alten indogermanischen Sprachen erschließen, sondern auch Textcorpora in "benachbarten" nicht-indogermanischen Sprachen, Texte in weniger verbreiteten modernen indogermanischen Sprachen, in kaukasischen Sprachen u.a.m. Der Einstieg erfolgt über die URL <http://titus.uni-frankfurt.de/texte/texte.htm>, der aktuelle Stand des verfügbaren Materials ist auf der Seite <http://titus.uni-frankfurt.de/texte/texte2.htm> dokumentiert.

## 2. Das TITUS-Retrievalsystem

Begünstigt durch die rasche Entwicklung der Webtechnik in jüngster Zeit ist es möglich geworden, über den einfachen Austausch von Textdateien (via Up- und Download) hinauszukommen und anstelle eines statischen Datenpools ein interaktives Retrieval-System zu entwickeln, das für die genannten Überlieferungen zahlreiche zusätzliche Informationsfelder erschließt. Dabei handelt es sich zunächst um eine tiefergehende Quellendokumentation, die die digitale graphische Erfassung der Überlieferungsträger, meist Handschriften oder Inschriften, umfasst. Zwei weitere Kernpunkte, an denen zur Zeit intensiv gearbeitet wird, betreffen die automatische Auswertung der Textdaten, nämlich einmal im Sinne einer textübergreifenden Eruierung und Ausgabe von Belegstellen (Konkordanzerstellung), zum anderen im Sinne einer linguistischen Analyse der in den Texten enthaltenen Wortformen. Die hierbei zu beachtenden sprach- und schriftspezifischen Probleme sowie die vom TITUS-Projekt eingeschlagenen Lösungswege seien an vier typischen Beispielen aufgezeigt.

### *2.1. Beispiel 1: Tocharisch*

Unter dem Namen "Tocharisch" bezieht man sich auf zwei eng miteinander verwandte, als Ost- und Westtocharisch oder "A-" und "B-Tocharisch" diffe-

renzierte indogermanische Sprachen, die etwa in der Zeit des 4.-8. Jh. n.Chr. in Chinesisch-Turkestan (der heutigen chinesischen Provinz Hsinking-Uygur) gesprochen und geschrieben wurden. Das uns verfügbare Material besteht aus ca. 4000 Handschriften, die im Zuge von drei Preußischen Expeditionen zu Beginn des 20.Jhs. rund um die Oase Turfan zutage gefördert wurden und heute von der Berlin-Brandenburgischen Akademie der Wissenschaften verwaltet werden (sog. Berliner Sammlung); ca. 1000 weitere Handschriften finden sich in Paris, London, St. Petersburg und an anderen Orten. Geschrieben sind die Dokumente in einer lokalen Abart der indischen Brāhmī-Schrift; sie enthalten zumeist aus dem Sanskrit oder anderen Sprachen übersetzte Textstücke buddhistischen Inhalts.

Im Rahmen des TITUS-Projekts, das in diesem Bereich seit 1995 intensiv mit der Stiftung Preußischer Kulturbesitz zusammenarbeitet, werden die tocharischen Materialien sukzessive in folgender Weise aufgearbeitet: Zunächst erfolgt eine Digitalisierung bereits in Druckform ediert vorliegender Textmaterialien, wobei die Texte wie in den gedruckten Editionen selbst in Umschrift eingegeben werden. Der zweite Schritt besteht in der Digitalisierung der Originalhandschriften, wobei aus Qualitätsgründen, aber auch zum Zwecke der Konservierung, der Umweg über die Erstellung von Farbdias gewählt wurde, die dann mit einem hochauflösenden Dia-Scanner eingelesen werden. Mit den so digitalisierten Handschriften werden die elektronischen Texte neu kollationiert und, wo erforderlich, überarbeitet. Das eigentlich Retrieval wird durch eine Vorindizierung der digitalen Texte vorbereitet, für die das WordCruncher Retrieval System eingesetzt wird.

Das erste Ergebnis dieser Bearbeitung ist die über <http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/thtframe.htm> verfügbare Ausgabe der edierten Teile der Berliner Sammlung (1100 Handschriften). Die auf HTML-Basis in Frames angelegte Ausgabe erschließt die Texte nach den Nummern des aktuellen Berliner Katalogs ("THT"), aber auch nach der Numerierung in den Druckausgaben (von Sieg/Siegling); gemeinsam mit den Texten, die in einer vereinfachten, in Unicode (UTF-8) codierten Umschrift gehalten sind, werden, soweit verfügbar, Bilder der betreffenden Handschriftenblätter bereitgestellt (ca. 350 Handschriften sind bisher nicht auffindbar gewesen; es handelt sich vermutlich um Kriegsverluste).

Für das eigentliche linguistische Retrieval (Suche nach Wörtern, Wortformen etc. und ihren Belegstellen) hat das TITUS-Projekt 1996 (mit Unterstützung der WordCruncher Company) einen WordCruncher Web-Server einrichten können, mit dem die umfangreichen Leistungsmerkmale dieses (von der Brigham Young University entwickelten) Retrievalsystems online verfügbar gemacht werden können (siehe <http://titus.fkidg1.uni-frankfurt.de/texte/tituswc2.htm>). Dies beinhaltet insbesondere die Möglichkeit einer sprachspezifischen Indizierung auch gemischtsprachiger Texte, die die Besonderheiten der jeweiligen Sprach- und Schriftstrukturen (z.B. alphabetische Anordnung,

Gleichwertigkeit unterschiedlicher Zeichen etc.) zu berücksichtigen gestattet. Die Nachteile des WordCruncher Web-Servers bestehen im wesentlichen darin, dass die Codierung der Texte nach wie vor auf einer 8-Bit-Basis erfolgen muss, was die Notwendigkeit sich überlappenden Spezialfonts mit sich bringt ("Font mapping"), und dass die Nutzbarkeit des Systems plattformabhängig ist, da es bisher lediglich für MS-Windows-Oberflächen verfügbar ist. Es kommt hinzu, dass für die Online-Kommunikation ein spezifisches Protokoll verwendet wird, was vielfach zu Übertragungsproblemen führt.

Da es zudem unsicher ist, ob und in welcher Form das WordCruncher-System weitergepflegt wird, wird im Rahmen des TITUS-Projekts derzeit ein eigenständiges Retrievalsystem entwickelt, das entsprechende Analysemöglichkeiten in einer HTTP-basierten Online-Bearbeitung bietet. Grundlage dieses Systems ist strukturierter Text auf HTML-Grundlage, der in Unicode codiert und, entsprechend dem jeweils üblichen Referenzierungssystem, in Einheiten (z.B. Text, Buch, Kapitel, Absatz, Satz oder Handschrift, Seite, Zeile) zerlegt ist. Im Laufe einer Vorindizierung werden sämtliche Wortformen, die in dem so aufbereiteten Text enthalten sind, mit ihrer Referenzangabe in einer Datenbank abgelegt; der Zugriff auf diesen Index kann dann durch manuelle Eingabe in einem Suchformular erfolgen (textbezogen: <http://titus.fkidl.uni-frankfurt.de/texte/etcc/toch/tocha/tochalex.htm>; textunabhängig: <http://titus.fkidl.uni-frankfurt.de/database/titusinx/titusinx.htm>), aber auch über Hyperlinks, mit denen jedes Wort im Text versehen ist. Ein Prototyp dieses Ausgabemodus ist für das A-tocharische Corpus über <http://titus.uni-frankfurt.de/texte/texte2.htm#toch> bzw. <http://titus.fkidl.uni-frankfurt.de/texte/etcc/toch/tocha/tocha.htm> verfügbar.

Durch die zweiteilige Grundkonzeption - Verknüpfung reiner Textdaten mit einer referentiellen Datenbank - lässt sich nicht nur der Aufwand des Taggings in den eigentlichen Textfiles auf ein Minimum reduzieren, was der Übertragungsgeschwindigkeit im Netz zugute kommt, sondern es können mit geringem Aufwand auch weitergehende Retrievalelemente hinzugegeben werden. So ist das System bereits jetzt in der Lage, beliebige tocharische Verbalformen, auch ungeachtet eventueller Schreibvariationen, korrekt zu bestimmen. So wird z.B. bei Aktivierung des betreffenden Hyperlinks durch Anklicken der Verbalformen nämseñc "sie verneigen sich" oder kumseñc "sie kommen" in dem A-tocharischen Text THT 634 (<http://titus.fkidl.uni-frankfurt.de/texte/etcc/toch/tocha/tocha001.htm>) auf dem TITUS-NT-Server ein ASP-Script aufgerufen (<http://titus.fkidl.uni-frankfurt.de/database/titusinx/titusinx.asp?LXLANG=58285&LXWORD=nE400mseF100cE400&LCPL=0&TCPL=0&C=H&PF=26>), das nicht nur sämtliche Belegstellen dieser Formen in den Texten aus der Datenbank extrahiert, sondern auch ihre Bestimmung als 3. Person Plural Präsens Aktiv; in der Ausgabeseite selbst ist weiter ein Hyperlink auf die betreffende Verbalwurzel (näm- "(sich) verneigen" bzw. käm- "kommen") enthalten, der seinerseits über einen ASP-Aufruf zur Ausgabe der kompletten Paradigmen

dieser Verben führt (<http://titus.fkidg1.uni-frankfurt.de/database/titusinx/tochverb.asp?R=kE400m2D00>). Es sei angemerkt, dass die Verwendung von Unicode bei der erforderlichen Übergabe von Parametern in den URL-Kommandozeilen heute noch nicht möglich ist, weshalb die Angaben in eine gemischte ASCII-Notierung umgesetzt werden müssen (z.B. kE400m2D00 für kām-).

## *2.2. Beispiel 2: Avestisch*

Während die am Tocharischen illustrierten Aufgaben und Lösungen für die meisten im Rahmen des TITUS-Projekts behandelten Sprachen ähnlich sind, bringen bestimmte Überlieferungen spezielle Probleme mit sich. Dies gilt z.B. für das Avestische, eine altiranische Sprache, in der die heiligen Texte der von Zarathustra begründeten zoroastrischen Religion niedergelegt sind, die etwa aus der Zeitspanne von 1000 - 300 v.Chr. stammen dürften, in handschriftlicher Form aber erst ab dem 13. Jh. n.Chr. greifbar werden. Eine Besonderheit besteht dabei nicht nur in der Verwendung einer eigens für das Avesta geschaffenen Schrift (eine phonetisch exakte, linksläufige Buchstabenschrift, die bisher noch nicht in Unicode codierbar ist), sondern auch darin, dass in den meisten der ca. 180 erhaltenen Handschriften avestische Textstücke im stetigen Wechsel mit mittelpersischen oder Sanskrit-Stücken angeordnet sind; dabei zeigen sich vielfach historische Abhängigkeiten zwischen einzelnen Handschriften oder -zweigen, hinter denen sich eine höchst komplexe stemmatische Struktur verbirgt. Hinzu kommt, dass die Texte selbst - ihrer Verwendung im liturgischen Zusammenhang entsprechend - in verschiedene, auch chronologisch differenzierbare Schichten zerfallen und immer wieder Repetitionen unterschiedlicher Art aufweisen, so dass für das Retrieval spezifische Deklarationen erforderlich sind.

Bei der Bearbeitung des Avesta-Corpus versucht das TITUS-Projekt, diesen Grundbedingungen gerecht zu werden, indem es nicht einfach den kritischen Text der bis heute maßgeblichen Ausgabe von K.F. Geldner reproduziert, sondern die verschiedenen Textschichten als solche herauszuarbeiten versucht, was auch eine Restituierung von (in den Handschriften und/oder der Druckausgabe meist abgekürzt notierten) Repetitionen umfasst (cf. <http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/avest.htm>). Für den Text der sog. Yasna-Liturgie wird darüber hinaus eine Handschriftenkonkordanz aufbereitet, die ein Retrieval der gesamten handschriftlichen Variation einer gegebenen Wortform ermöglichen soll; ein Prototyp ist bereits abrufbar (cf. z.B. <http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasna044.htm>: durch Anklicken des mit der Wortform dĀ□tĀ□ in Y. 43,13c verknüpften Hyperlinks wird wiederum ein ASP-Script aufgerufen, das nicht nur sämtliche zehn möglichen Bestimmungen dieser Wortform aus der Datenbank extrahiert, sondern auch die handschriftlichen Varianten, die für sie verfügbar sind (<http://titus.fkidg1>.

### 2.3. Beispiel 3: Vedisch

Wieder andere Spezialprobleme ergeben sich beim Vedischen, der ältesten Ausprägung des Altindischen oder Sanskrit. Hinsichtlich ihrer Überlieferungsform zeigen die vedischen Texte zwar viele Übereinstimmungen mit dem Avestischen, sowohl inhaltlich (es handelt sich ebenfalls durchweg um religiöse Texte) als auch äußerlich (die Texte stammen etwa aus derselben Epoche, ca. 1500 - 300 v.Chr., und sind ebenfalls erst sehr viel später schriftlich niedergelegt worden). Die besonderen Probleme ergeben sich jedoch vor allem dadurch, dass entsprechend der indischen Schrifttradition Wortgrenzen in den Originaldokumenten normalerweise nicht als solche notiert sind, wobei zusätzlich starke lautliche Veränderungen im Wortan- und -auslaut entstehen, wo Wörter im Satzzusammenhang aufeinanderfolgen (Vokalelisionen, -kontraktionen etc.; sog. "Sandhi").

Der Umgang mit diesem Problem hat schon in der klassischen Editionspraxis zu unterschiedlichen Lösungsansätzen geführt, wobei die Abbildung von Wortgrenzen gemeinhin leichter durchzuführen war, wenn man anstelle der Originalschrift (meist der Devanāgarī-Schrift) eine Latein-Umschrift wählte. Da die auf dem TITUS-Server verfügbaren elektronischen Fassungen der vedischen Texte (derzeit ca. 75% des gesamten vedischen Corpus) ihrerseits meist auf gedruckten Ausgaben beruhen, spiegeln sie die unterschiedlichen Gepflogenheiten derzeit noch weitgehend wieder, so dass ein einheitliches Retrieval für jede Wortform ungeachtet ihrer umgebungsbedingten formalen Variation noch nicht gewährleistet werden kann. Die Bedingung für ein solches Retrieval wäre die - in der indischen Tradition selbst verankerte - Aufarbeitung der Texte in einer Erfassungsvariante, bei der jede Wortform ungeachtet ihres Kontexts in einer Normalform eingesetzt wird; ein solcher "Pada-Pāṇi-ha" (Pada-Patha) ist bisher aber nur für wenige der Texte verfügbar (cf. z.B. die TITUS-Version des bedeutendsten Texts der Sammlung, der R̥gveda-Saá'fhitā, (Rgveda-Samhita) bei der der undifferenzierte Ausgangstext, der sog. "Saá'fhitā-Pāṇi-ha" (Samhita-Patha), und der Pada-Pāṇi-ha (Pada-Patha) einander interlinear gegenübergestellt sind: <http://titus.uni-frankfurt.de/texte/etcs/ind/aind/ved/rv/mt/rv.htm>). In der letztlich angestrebten Lösung sollen die Pada-Pāṇi-ha-(Pada-Patha)-Versionen gewissermaßen unsichtbar in den Hintergrund treten und lediglich zur eindeutigen Referenzierung der im fließenden Text enthaltenen Wortformen dienen. Ein besonderes Problem wird sich dabei einstellen, wenn, wie beabsichtigt, die Texte insgesamt auch in Unicode-codierter Devanāgarī-Schrift aufbereitet werden sollen (vgl. den Prototyp für die R̥gveda-Saá'fhitā (Rgveda-Samhita) unter <http://titus.uni-frankfurt.de/texte/etcd/ind/aind/ved/rv/mt/rv.htm>); denn die korrekte Positio-

nierung der Hyperlinks wird es in diesem Fall erforderlich machen, Tags mitten in ein gegebenes "Silbenzeichen" ("Aká'fara" - Aksara) hineinzusetzen, was jedoch einen Verstoß gegen Unicode-Konventionen (und dementsprechend unvorhersagbare Resultate bei der Browser-Wiedergabe) bedeuten kann. Es sei angemerkt, dass die Verwendung von Unicode-codierter Devanāgarī auf Webseiten zur Zeit ohnehin noch vielfältige Probleme mit sich bringt, insbesondere bei vedischen Texten, bei denen das Zeicheninventar um verschiedene Akzentzeichen ergänzt ist.

#### *2.4. Beispiel 4: Neupersisch*

Dass die Verwendung von Originalschriften trotz des Einsatzes von Unicode bei der Aufbereitung der Textmaterialien spezielle Probleme aufwerfen kann, gilt nicht nur für den indischen Bereich. Im Rahmen des TITUS-Projekts betrifft dies z.B. auch Materialien in neupersischer Sprache, wobei für das Projekt im wesentlichen das "klassische" Neupersisch mit Texten aus etwa dem 8. - 13. Jh. n.Chr. relevant ist. Seit dieser Zeit wird das Persische in arabischer Schrift geschrieben, die sich von der Lateinschrift nicht nur durch die umgekehrte Schreibrichtung (linksläufig) unterscheidet. Auch hier ist es im wesentlichen die Einsetzung von (notwendigerweise ASCII-basierten) Hyperlink-Tags im arabischen Kontext, die spezielle Anforderungen an die die HTML-Seiten wiedergebenden Browser stellt. Auf dem TITUS-Server werden neupersische Texte deshalb zumindest für eine gewisse Zeit neben der originalschriftlichen Fassung noch in lateinischer Umschrift angeboten (vgl. z.B. die beiden Varianten des Epos von Vāčs und Rāmān unter <http://titus.uni-frankfurt.de/texte/etca/iran/niran/npers/visrpx/visrp.htm> und <http://titus.uni-frankfurt.de/texte/etca/iran/niran/npers/visrp/visrp.htm>).

Eine zweite Problematik, die durch die arabische Schrift bedingt ist, betrifft die Lautgestalt der Wortformen. Da die arabischen Buchstaben im großen und ganzen nur das Konsonantengerippe einer gegebenen Wortform darstellen, wird eine automatische Umsetzung von der arabischen in eine Lateingraphie zunächst, wie im o.g. Text, nur eine fremd anmutende transliterative Darstellung ermöglichen. Für ein linguistisches Retrieval wäre es aber zumindest wünschenswert, wenn die tatsächliche Lautung der Wörter, unter Einschluss ihrer Vokale, die Grundlage bilden würde. Diese Aufgabe soll in Zukunft wiederum über eine verknüpfte Datenbank gelöst werden.

### **3. Zusammenfassung und Ausblick**

Die Grundlagen des TITUS-Retrievalsystems, das derzeit noch in der Erprobungsphase steckt, seien abschließend noch einmal zusammengestellt. Sie umfassen:



- Aufarbeitung der Texte in Unicode-Codierung mit einheitlicher HTML-Struktur (über Style-Sheet-Vorgaben);
- eindeutige Sprachzuweisung für alle in den Texten enthaltenen Wortformen;
- Vorindizierung des Textmaterials mit exakter Stellenreferenzierung;
- Erfassung der auf die Wortformen bezogenen Daten in einer Datenbank (wegen seiner Unicode-Kompatibilität wird derzeit MS-Access verwendet), wobei durch Rückgriff auf sprachspezifische Vorgaben Schreibvarianten etc. einander zugeordnet werden;
- SQL-basierter Zugriff auf die Datenbank via ASP-Seiten (Windows-NT-Server), die über Javascript-basierte Hyperlinks direkt aus dem HTML-Text oder über Eingabeformulare aufgerufen werden. Im letzteren Falle können übrigens auch Wildcards benutzt werden (\* für beliebige Zeichenfolgen, ? für beliebige Einzelzeichen).

An folgenden Weiterentwicklungen wird derzeit gearbeitet:

- Ausweitung der Suchmöglichkeiten (verbundene Suche mehrerer Wortformen im Kontext etc.; sprachspezifische Wildcards, z.B. für "alle Vokale" etc.);
- Ausweitung der Wiedergabe in Originalschriften (Voraussetzungen: Unicode-Ergänzung, z.B. für Avestisch, Mittelpersisch etc.; Ausbau vorhandener Unicode-Implementierungen, z.B. für Syrisch: vgl. <http://titus.fkidg1.uni-frankfurt.de/texte/etcc/semit/syr/syrbible/syrbi.htm>).
- Geprüft wird der Übergang von HTML-Strukturen zu XML-Strukturen.